Stopped Markov Decision Processes with Multiple Constraints (Perspective and problem for Dynamic Programming with uncertainty)

AUTHOR(S):

Horiguchi, Masayuki

CITATION:

Horiguchi, Masayuki. Stopped Markov Decision Processes with Multiple Constraints (Perspective and problem for Dynamic Programming with uncertainty). 数理解析研究所講究録 2001, 1207: 41-54

ISSUE DATE:

2001-05

URL:

http://hdl.handle.net/2433/41042

数理解析研究所講究録 1207 巻 2001 年 41-54

Stopped Markov Decision Processes with Multiple Constraints

千葉大学・自然科学研究科　堀口正之 (Masayuki HORIGUCHI)

Abstract

In this paper, a optimization problem for stopped Markov decision processes with vector-valued terminal reward and multiple running cost constraints is considered. Applying the idea of occupation measures and using the scalarization technique for vector maximization problems we obtain the equivalent Mathematical Programming problem and show the existence of a Pareto optimal pair of stationary policy and stopping time requiring randomization in at most k states, where k is the number of constraints. Moreover Lagrange multiplier approaches are considered. The saddle-point statements are given, whose results are applied to obtain a related parametric Mathematical Programming, by which the problem is solved. Numerical examples are given.

Key words: Stopped Markov decision process, multi-objective, multiple constraints, Mathematical Programming formulation, Lagrange multiplier.

1 Introduction

The aim of this paper is to establish a Mathematical Programming method for finite state stopped Markov decision processes(MDPs) with vector-valued terminal reward and multiple running cost constraints. In the preceding paper Horiguchi [9], we consider a optimization problem for stopped Markov decision processes with a constrained stopping time. The problem is solved through randomization of stopping times and Mathematical Programming formulation by occupation measures. Here, we consider the vector-valued and multiple constrained case. The optimality is defined by the concept of efficiency, based on a pseudo-order preference relation $\preccurlyeq_K$ induced by a closed convex cone K in $\mathbb{R}^p$. Then a Pareto optimization with respect to the pseudo-order $\preccurlyeq_K$ is considered(cf. [6, 18]). Applying the idea of Horiguchi[9], we derive a related Mathematical Programming to solve the problem treated in this paper and show the existence of a Pareto optimal pair of stationary policy and stopping time requiring randomization in at most k states, where k is the number of constraints. Also, introducing a corresponding Lagrange function, the saddle-point statements for our constrained problem are given, whose results are applied to obtain a related parametric Mathematical Programming, by which the problem is solved. Numerical examples are given to illustrate the results. For the unconstrained case, refer to Furukawa and Iwamoto[7], Hordijk[8] and Rieder[16]. A Lagrangian approach to a constrained optimal stopping problem has been discussed originally by Nachman[15] and Kennedy[12]. For the dynamic programming approach to the constrained Markov decision processes, see Horiguchi et al[10].

In the reminder of this section, we shall establish some notation that will be used throughout this paper referring to the preceding work[9] and define the vector-valued optimization problem with multiple constraints. Also, a Pareto optimal pair of policy and randomized stopping time is defined.

Let S and A be finite sets denoted by $S = \{1, 2, \dots N\}$ and $A = \{1, 2, \dots, K\}$. The stopped Markov decision model consists of five objects:

$$(S, A, \{p_{ij}(a) : i, j \in S, a \in A\}, \{c^l, l = 1, 2, \dots, k\}, \boldsymbol{r})$$

where S and A denote the state and action spaces respectively and $\{p_{ij}(a)\}$ is the law of motion, i.e., for each $(i, a) \in S \times A, p_{ij}(a) \geqq 0$ and $\sum_{j \in S} p_{ij}(a) = 1$, and $c^l = c^l(i, a)$, $l = 1, 2, \dots, k$, are running cost functions on $S \times A$, which will be related to k constraints, and $\boldsymbol{r} = \boldsymbol{r}(i) = (r^1(i), \dots, r^p(i))$ is a vector-valued terminal reward function on S when selecting "stop" in state i.

When the system is in state $i \in S$, if we select "stop" the process terminates with the terminal reward $\boldsymbol{r}(i)$. If we select "continue" and take $a \in A$, we move to a new state $j \in S$ selected according to the probability distribution $p_{i\cdot}(a)$ and the costs $c^l(i, a), l = 1, 2, \dots, k$ are incurred. This process is repeated from the new state $j \in S$. Let x_t, a_t be the state and action at time t and $h_t = (x_1, a_1, \dots, x_t) \in (S \times A)^{t-1} \times S$ the history up to time $t (t \geqq 1)$. A policy for controlling the system is a sequence $\pi = (\pi_1, \pi_2, \dots)$ such that, for each $t \geqq 1$, π_t is a conditional probability measure on A given history h_t with $\pi_t(A|(x_1, a_1, \dots, x_t)) = 1$ for each $(x_1, a_1, \dots, x_t) \in (S \times A)^{t-1} \times S$. Let Π denotes the set of all policies. A policy $\pi = (\pi_1, \pi_2, \dots)$ is a Markov policy if π_t is a function of only x_t, i.e., $\pi_t(\cdot|x_1, a_1, \dots, x_t) = \pi_t(\cdot|x_t)$ for all $(x_1, a_1, \dots, x_t) \in (S \times A)^{t-1} \times S$. A Markov policy $\pi = (\pi_1, \pi_2, \dots)$ is stationary if there exists a conditional probability on A, $w(\cdot|i)$, given $i \in S$ such that $\pi_t(\cdot|x_t) = w(\cdot|x_t)$ for all $x_t \in S$ and $t \geqq 1$, and denoted simply by w. A stationary policy w is called deterministic if there exists a map $g : S \to A$ with $w(g(i)|i) = 1$ for all $i \in S$ and such a policy is identified by g. The sets of all Markov, stationary and deterministic policies will be denoted by Π_M, Π_S and Π_D respectively. Note that $\Pi_D \subset \Pi_S \subset \Pi_M \subset \Pi$. The sample spaces is the product space $\Omega = (S \times A)^\infty$. Let X_t, Δ_t be random quantities such that $X_t(\omega) = x_t$ and $\Delta_t(\omega) = a_t$ for all $\omega = (x_1, a_1, x_2, a_2, \dots) \in \Omega$. For any given policy $\pi \in \Pi$ and initial distribution β on S we can specify the probability measure $\mathbb{P}^\pi_\beta$ on Ω in a usual way.

Let $H_t = (X_1, \Delta_1, \dots, X_t)$. Let $\mathcal{F}_t = \mathcal{B}(H_t), t \geqq 1$, where $\mathcal{B}(H_t)$ is the σ-field induced by H_t and $\mathcal{F}_\infty$ the smallest σ-field containing each $\mathcal{F}_t, t \geqq 1$. Let $\overline{N} = \{1, 2, \dots\} \cup \{\infty\}$. In order to solve our constrained decision problem described in the sequel, we need to introduce randomized stopping time (cf. [2, 5, 12])). To this purpose, enlarging Ω to $\overline{\Omega} := \Omega \times [0, 1]$, we can embed $(\Omega, \mathcal{F}_\infty)$ to $(\overline{\Omega}, \mathcal{F}_\infty \times \mathbb{B}_1)$, where $\mathbb{B}_1$ is Borel subsets of $[0, 1]$. For a filtration $\mathcal{F}^* = \{\mathcal{F}^*_t, t \in \overline{N}\}$ with $\mathcal{F}^*_t = \mathcal{F}_t \times \mathbb{B}_1$ we can assume without loss of generality that for each $t \in \overline{N}$

$$\mathcal{F}_t \subset \mathcal{F}^*_t. \tag{1.1}$$

We call a map $\overline{\tau} : \overline{\Omega} \to \overline{N}$ a randomized stopping time (hereafter called RST) w.r.t. $\mathcal{F}^*$ if $\{\overline{\tau} = t\} \in \mathcal{F}^*_t$ for each $t \in \overline{N}$. For simplicity, the upper bar of RST $\overline{\tau}$ will be omitted and written by τ with some abuse of notation. The class of RSTs w.r.t. $\mathcal{F}^*$ will be denoted by $\mathcal{S}$. For each initial distribution β and each policy $\pi \in \Pi$, we denote the probability measure on $\overline{\Omega}$ by $\overline{\mathbb{P}}^\pi_\beta$, where $\overline{\mathbb{P}}^\pi_\beta = \mathbb{P}^\pi_\beta \times \lambda$ and λ is Lebesgue measure on $\mathbb{B}_1$.

Let $\mathbb{R}^p$ be the set of real p-dimensional row vectors and $K \subset \mathbb{R}^p$ a nontrivial closed and pointed convex cone(cf. [17]). We introduce a pseudo-order relation $\preccurlyeq_K$ on $\mathbb{R}^p$ by $x \preccurlyeq_K y$ iff $y - x \in K$. For a nonempty subset $U \subset \mathbb{R}^p$, a point $x \in U$ is called efficient with respect to the order $\preccurlyeq_K$ on $\mathbb{R}^p$ if $x \preccurlyeq_K y$ for some $y \in U$ implies $x = y$. Let $\boldsymbol{e}(U)$ denote the set of all efficient points of U with respect to $\preccurlyeq_K$.

For any $\alpha = (\alpha^1, \ldots, \alpha^k) \in \mathbb{R}^k$ and initial distribution β on S, let

$$\Lambda^k(\alpha, \beta) := \{(\pi, \tau) \in \Pi \times \mathcal{S} \mid \overline{\mathbb{E}}^{\pi}_{\beta} \sum_{t=1}^{\tau-1} c^l(X_t, \Delta_t) \leqq \alpha^l \text{ for } l = 1, 2, \ldots, k\} \tag{1.2}$$

where $\overline{\mathbb{E}}^{\pi}_{\beta}$ is the expectation with respect to $\overline{\mathbb{P}}^{\pi}_{\beta}$.

We shall consider the vector-valued constrained optimization problem(**VCOP**):

$$\begin{aligned} \textbf{VCOP}: \ & \text{Maximize } \overline{\mathbb{E}}^{\pi}_{\beta} \boldsymbol{r}(X_\tau) := (\overline{\mathbb{E}}^{\pi}_{\beta} r^1(X_\tau), \ldots, \overline{\mathbb{E}}^{\pi}_{\beta} r^p(X_\tau)) \\ & \text{subject to } (\pi, \tau) \in \Lambda^k(\alpha, \beta). \end{aligned}$$

A pair $(\pi^*, \tau^*) \in \Lambda^k(\alpha, \beta)$ is called Pareto optimal if

$$\overline{\mathbb{E}}^{\pi^*}_{\beta} \boldsymbol{r}(X_{\tau^*}) \in \boldsymbol{e}(\{\overline{\mathbb{E}}^{\pi}_{\beta} \boldsymbol{r}(X_\tau) | (\pi, \tau) \in \Lambda^k(\alpha, \beta)\}). \tag{1.3}$$

Note that if $c^l \equiv 1$ for $l = 1, 2, \ldots, k$, the running cost constraints are reduced to $\overline{\mathbb{E}}^{\pi}_{\beta} \tau \leqq d$, where $d = \min_{1 \leq l \leq k} \alpha^l + 1$, whose case have been studied in [9], so that works in this paper are thought of as a generalization of those in [9].

Now, we define F-representation of a RST (cf. [9, 11]). For any RST $\tau \in \mathcal{S}$ and $t \in \overline{N}$, let

$$g_t(\omega) := \lambda(\{\tau = t\}_\omega) \quad (\omega \in \Omega),$$

where $\{\tau = t\}_\omega$ is the ω-section defined by $\{\tau = t\}_\omega = \{x \in [0,1] | (\omega, x) \in \{\tau = t\}\}$. Note that g_t is $\mathcal{F}_t$-measurable for $t \geqq 1$. From this g_t $(t \in \overline{N})$, we define the set $f = (f_t)_{t \in \overline{N}}$ as follows:

$$f_t := \frac{g_t}{1 - \sum_{k=1}^{t-1} g_k}, \quad t \in \overline{N} \tag{1.4}$$

where if the denominator is 0 in (2.1) let $f_t = 1$.

Let $F = \{a = (a_j)_{j \in \overline{N}} : 0 \leqq a_j \leqq 1, a_\infty = 1, \text{ and if } a_j = 1 \ a_i = 1 \text{ for } i > j\}$. Then, it is shown in Lemma 2.1 of [9] that $f : \Omega \to F$, each f_t is $\mathcal{F}_t$-measurable and for any initial distribution β and pair $(\pi, \tau) \in \Pi \times \mathcal{S}$ it holds

$$f_t = \frac{\overline{\mathbb{P}}^{\pi}_{\beta}(\tau = t | H_t)}{\overline{\mathbb{P}}^{\pi}_{\beta}(\tau \geqq t | H_t)}, \ \mathbb{P}^{\pi}_{\beta}\text{-a.s.} \tag{1.5}$$

The set $f = (f_t)_{t \in \overline{N}}$ constructed from $\tau \in \mathcal{S}$ is called F-representation of τ, denoted by $f^\tau = (f^\tau_t)_{t \in \overline{N}}$.

Conversely, letting $f = (f_t)_{t \in \overline{N}}$ be any function $f : \Omega \to F$ such that for each $t \in \overline{N}$ f_t is $\mathcal{F}_t$-measurable, we can construct the RST from f (see Lemma 2.2 of [9]), denoted by τ^f. Hence, there is one-to-one correspondence between $\mathcal{S}$ and the set of F-representations $f = (f_t)_{t \in \overline{N}}$. Using this fact, we define several types of RSTs.

Since the corresponding F-representation f^τ_t is $\mathcal{F}_t$-measurable $(t \geqq 1)$, f^τ_t is a function of $H_t = (X_1, \Delta_1, \ldots, X_t)$.

Definition 1.1. *If f_t^τ is depending only on X_t, that is, $f_t^\tau(H_t) = f_t^\tau(X_t)$ for all $t \geqq 1$, the RST τ is called Markov. A Markov RST is called stationary if there exists a function $\delta : S \to [0,1]$ such that $f_t^\tau(X_t) = \delta(X_t)$ for all $t \geqq 1$, denoted simply by δ. When $\delta(i) \in \{0,1\}$ for all $i \in S$, the stationary RST δ is called deterministic.*

We denote the sets of all Markov RSTs, all stationary RSTs and all deterministic RSTs by $\mathcal{S}_M, \mathcal{S}_S$ and $\mathcal{S}_D$ respectively. With some abuse of notation, we often use for a stationary RST $\tau \in \mathcal{S}_S$ the F-representation δ with $\delta = (f_t^\tau)$ such as $\delta \in \mathcal{S}_S$.

It is easily shown(cf. [9]) that the running costs and terminal reward under a $\tau \in \mathcal{S}$ and $\pi \in \Pi$ are represented using F-representation $f^\tau = (f_t^\tau)_{t \in \overline{N}}$ as follows.

$$\begin{aligned} &\overline{\mathbb{E}}_\beta^\pi \sum_{t=1}^{\tau-1} c^l(X_t, \Delta_t) \qquad (1.6) \\ &= \sum_{t=1}^{\infty} \mathbb{E}_\beta^\pi (1-f_1)(1-f_2)\cdots(1-f_{t-1}) f_t \Big(\sum_{n=1}^{t-1} c^l(X_n, \Delta_n)\Big), \quad l = 1, 2, \ldots, k \end{aligned}$$

and

$$\begin{aligned} &\overline{\mathbb{E}}_\beta^\pi r^m(X_\tau) \qquad (1.7) \\ &= \sum_{t=1}^{\infty} \mathbb{E}_\beta^\pi (1-f_1)(1-f_2)\cdots(1-f_{t-1}) f_t r^m(X_t, \Delta_t), \quad m = 1, 2, \ldots, p. \end{aligned}$$

Let K^* denote the dual cone of a convex cone $K \subset \mathbb{R}^p$, i.e., $K^* = \{b \in \mathbb{R}^p : \langle b, x \rangle \geq 0 \text{ for all } x \in K\}$ where $\langle \cdot, \cdot \rangle$ means inner product in $\mathbb{R}^p$.

The following result is well known(cf. [3])

Lemma 1.1. *Let $B \subset \mathbb{R}^p$ be compact and convex set. Then $x \in \boldsymbol{e}(B)$ if and only if there exists $b \in K^*(b \neq 0)$ such that $\langle b, x \rangle \geq \langle b, y \rangle$ for all $y \in B$.*

In Section 2, the running and stopped occupation measures are introduced, by which **VCOP** is reduced equivalently to multi-objective Mathematical Programming problem. In Section 3, studying the properties of the set of running occupation measures we can prove the existence of a Pareto optimal pair of stationary policy and stopping time requiring randomization in at most k states. Section 4 is devoted to Lagrange approaches for **VCOP**. Finally, the proof of key Lemma used in Section 3 is given in Section 5.

2 Occupation measures in Stopped MDPs

In this section, we introduce two types of occupation measures in stopped MDPs and consider the properties of them.

Definition 2.1. *For any initial distribution β and a pair (π, τ) with $\overline{\mathbb{E}}_\beta^\pi[\tau] < \infty$, we define the measure $x(\beta, \pi, \tau)$ on $S \times A$, called the running occupation measure, by*

$$x(\beta, \pi, \tau; i, a) := \sum_{t=1}^{\infty} \overline{\mathbb{P}}_\beta^\pi (X_t = i, \Delta_t = a, \tau > t) \quad \text{for } i \in S, a \in A. \qquad (2.1)$$

Definition 2.2. *For any initial distribution β and a pair (π,τ) with $\overline{\mathbb{E}}^{\pi}_{\beta}[\tau] < \infty$, we define the measure $y(\beta,\pi,\tau)$ on $S \times A$, called the stopped occupation measure, by*

$$y(\beta,\pi,\tau;i,a) := \sum_{t=1}^{\infty} \overline{\mathbb{P}}^{\pi}_{\beta}(X_t = i, \Delta_t = a, \tau = t) \quad \text{for } i \in S, a \in A. \tag{2.2}$$

The state running and stopped occupation measures will be defined by

$$x(\beta,\pi,\tau;i) := \sum_{a\in A} x(\beta,\pi,\tau;i,a) \quad \text{and}$$
$$y(\beta,\pi,\tau;i) := \sum_{a\in A} y(\beta,\pi,\tau;i,a) \quad \text{for all } i \in S.$$

For any $\delta : S \to [0,1]$ and conditional distribution $w(\cdot|i)$ on A given $i \in S$, we define by $P^{\delta}(w)$ the $N \times N$ matrix where (i,j)th element is $P_{ij}(w)(1-\delta(j)) := \sum_{a\in A} p_{ij}(a)w(a|i)(1-\delta(j))$ or simply $(P^{\delta}(w))_{ij}$. With some abuse of notation, for any initial distribution β and $(\pi,\tau) \in \Pi \times \mathcal{S}$, the row vector $x(\beta,\pi,\tau) \in \mathbb{R}^N$ is defined by

$$x(\beta,\pi,\tau) := (x(\beta,\pi,\tau;1),\ldots,x(\beta,\pi,\tau;N)).$$

Then, in the following lemma, the state stopped occupation measure is proved to be represented by the running one and for each $(w,\tau) \in \Pi_S \times \mathcal{S}_S$, the state running occupation measure is determined uniquely as a solution of the corresponding equations below.

Lemma 2.1. ([9]) *For any initial distribution β and pair $(\pi,\tau) \in \Pi \times \mathcal{S}$ with $\overline{\mathbb{E}}^{\pi}_{\beta}[\tau] < \infty$ we have the following:*

(i) $x(\beta,\pi,\tau;i) < \infty$ *and* $y(\beta,\pi,\tau;i) < \infty$ *for all* $i \in S$.

(ii) $\overline{\mathbb{E}}^{\pi}_{\beta}[\tau] = \sum_{i\in S} x(\beta,\pi,\tau;i) + 1$.

(iii) $y(\beta,\pi,\tau;i) = \beta(i) + \sum_{j\in S,a\in A} x(\beta,\pi,\tau;j,a)p_{ji}(a) - x(\beta,\pi,\tau;i)$.

Moreover, if $(w,\tau) \in \Pi_S \times \mathcal{S}_S$, then the state running occupation measure $x(\beta,w,\tau)$ is the unique solution to

$$x = \beta(1-\delta) + xP^{\delta}(w), \; x \in \mathbb{R}^N \tag{2.3}$$

where $\beta(1-\delta)$ is in $\mathbb{R}^N$ whose i-th component is $\beta(i)(1-\delta(i))$ and $\delta := f^{\tau} : S \to [0,1]$ is F-representation of τ.

Let $\mathbb{R}^{N\times K}$ be the set of real $N \times K$ matrices. For any subset $U \subset \Pi \times \mathcal{S}$, denote

$$\mathbb{X}^k(U) := \{x(\beta,\pi,\tau;i,a)_{i\in S,a\in A} : (\pi,\tau) \in U \cap \Lambda^k(\alpha,\beta)\}. \tag{2.4}$$

Note that $\mathbb{X}^k(U) \subset \mathbb{R}^{N\times K}$.

Here, we define the multi-objective Mathematical Programming problem(**MMP(I)**) related to **VCOP** as follows:

$$\begin{aligned}
\textbf{MMP(I)}: \text{ Maximize } & \sum_{i\in S} \boldsymbol{r}(i)y(i) := (\sum_{i\in S} r^1(i)y(i), \ldots, \sum_{i\in S} r^p(i)y(i)) \\
\text{subject to } & x \in \mathbb{X}^k(\Pi \times \mathcal{S}),\ y \in \mathbb{R}^N \text{ and} \\
& y(i) = \beta(i) + \sum_{j\in S, a\in A} x(j,a)p_{ji}(a) - x(i), \quad i \in S, \\
& \text{where } x(i) = \sum_{a\in A} x(i,a).
\end{aligned}$$

Then, we have the following theorem, which is proved from Lemma 1.1 by the use of (Theorem 3.1 of [9]).

Theorem 2.1. VCOP *is equivalent to* **MMP(I)**, *i.e., a pair* (π^*, τ^*) *is Pareto optimal for* **VCOP** *if and only if the corresponding* $\{x(\beta, \pi^*, \tau^*; i, a)\} \in \mathbb{X}^k(\Pi \times \mathcal{S})$ *is Pareto optimal for* **MMP(I)**.

Proof. From Lemma 1.1, an efficient point for **VCOP** is given by solving the following maximization problem for some $b \in K^*$:

$$\begin{aligned}
& \text{Maximize } \langle b, \overline{\mathbb{E}}^{\pi}_{\beta}\boldsymbol{r}(X_\tau)\rangle \\
& \text{subject to } (\pi, \tau) \in \Lambda^k(\alpha, \beta).
\end{aligned} \tag{2.5}$$

Applying (Theorem 3.1 of [9]) will complete the proof of Theorem 2.1. ∎

3 Mathematical Programming approach and Pareto optimal pair

In this section, we present another Mathematical Programming formulation by which **VCOP** is explicitly solved.

To this end, we define several basic sets below. For simplicity, we put $(x_{ia}) = \{x_{ia}\}_{i\in S, a\in A} \in \mathbb{R}^{N\times K}$ and $\delta = \{\delta(i)\}_{i\in S} \in \mathbb{R}^N$. For any initial distribution β on S and $\alpha = (\alpha^1, \ldots, \alpha^k) \in \mathbb{R}^k$, let

$$\hat{\mathbb{Q}}^k := \left\{ \begin{array}{l} ((x_{ia}), \delta) \in \mathbb{R}^{N\times K} \times \mathbb{R}^N : \\ \quad \text{(i)} \ \sum_{a\in A} x_{ia} = \beta(i)(1-\delta(i)) + \sum_{j\in S, a\in A} x_{ja}p_{ji}(a)(1-\delta(i)), \ (i \in S) \\ \quad \text{(ii)} \ 0 \leqq \delta(i) \leqq 1, \quad (i \in S) \\ \quad \text{(iii)} \ \sum_{i\in S, a\in A} c^l(i,a)x_{ia} \leqq \alpha^l, \quad (l = 1, 2, \ldots, k) \\ \quad \text{(iv)} \ x_{ia} \geqq 0, \quad (i \in S, a \in A) \end{array} \right\}, \tag{3.1}$$

$$\mathbb{Q}^k := \left\{ (x_{ia}) \in \mathbb{R}^{N\times K} : ((x_{ia}), \delta) \in \hat{\mathbb{Q}}^k \text{ for some } \delta \right\}. \tag{3.2}$$

We introduce the following assumption.

Assumption $(*)$. *For any* $w \in \Pi_S$ *and* $l(1 \leqq l \leqq k)$,

$$\max_{1 \leqq l \leqq k} c^l(i|w) > 0 \quad \text{for each} \quad i \in S \tag{3.3}$$

where $c^l(i|w) = \sum_{a \in A} c^l(i,a) w(a|i)$.

We have the following theorem, whose proof is similar to (Theorem 4.1, Lemma 4.1 and Theorem 4.2 in [9]) and omitted.

Theorem 3.1. *Suppose that Assumption* $(*)$ *holds. Then*

(i) $\mathbb{X}^k(\Pi \times \mathcal{S}) = \mathbb{X}^k(\Pi_M \times \mathcal{S}_M) = \mathbb{X}^k(\Pi_S \times \mathcal{S}_S)$.

(ii) $\mathbb{Q}^k = \mathbb{X}^k(\Pi_S \times \mathcal{S}_S)$.

(iii) $\mathbb{Q}^k$ *is compact and convex.*

The following corollary holds clearly from Theorem 3.1 and observing (3.2).

Corollary 3.1. $\mathbb{X}^k(\Pi_S \times \mathcal{S}_S)$ *is compact and convex.*

Remark. For any $((x_{ia}), \delta) \in \hat{\mathbb{Q}}^k$, we define a stationary policy w as follows:
For each $a \in A$ and $i \in S$,

$$w(a|i) = \begin{cases} \dfrac{x_{ia}}{x_i}, & \text{if } x_i > 0, \\ \text{any probability distribution on } A, & \text{if } x_i = 0, \end{cases} \tag{3.4}$$

where $x_i = \sum_{i \in A} x_{ia}$. Then, $x = (x_i)$ with $x_i = x(\beta, w, \delta; i), i \in S$ is given as a unique solution of (2.3).

Also, (i) and (iii) in (3.1) are rewritten as follows:

$$\begin{cases} (\mathrm{i}') & x_i = \beta(i)(1 - \delta(i)) + \sum_{j \in S} x_j P_{ji}(w)(1 - \delta(i)), i \in S \\ (\mathrm{iii}') & \sum_{i \in S} c^l(i,w) x_i \leqq \alpha^l, l = 1, 2, \ldots, k \end{cases} \tag{3.5}$$

where $c^l(i,w) = \sum_{a \in A} c^l(i,a) w(a|i)$.
Now, we define another multi-objective Mathematical Programming problem (**MMP(II)**) for **VCOP**:

$$\textbf{MMP(II)} : \text{Maximize} \quad \sum_{i \in S} \boldsymbol{r}(i) y_i$$

$$\text{subject to} \quad (x_{ia}) \in \mathbb{Q}^k, \quad y_i = \beta(i) + \sum_{j \in S, a \in A} x_{ja} p_{ji}(a) - \sum_{a \in S} x_{ia}, \quad i \in S$$

Here we get the following corollary which is obviously given from Theorem 2.1 and 3.1 and Corollary 3.1.

Corollary 3.2. *It holds that*

(i) **VCOP** *and* **MMP(II)** *are equivalent.*

(ii) *A Pareto optimal pair exists on* $\Pi_S \times \mathcal{S}_S$.

For any stationary policy $w \in \Pi_S$, let $n(w)$ be the total number of randomization under w, that is,

$$n(w) = \sum_{i \in S}(m(i,w) - 1),$$

where $m(i,w)$ is the number of elements in $\{a \in A | w(a|i) > 0\}$. Define

$$\Pi_S^k := \{w \in \Pi_S : n(w) \leqq k\},$$

and

$$\mathcal{S}_S^k := \{\tau \in \mathcal{S}_S | f^\tau(i) \in \{0,1\} \text{ except at most } k \text{ states}\}.$$

For $(x_{ia}) \in \mathbb{Q}^k$, $\mathcal{I}((x_{ia})) \subset \{1,2,\ldots k\}$ is defined as follows:

$$\mathcal{I}((x_{ia})) := \{l \in \{1,2,\ldots,k\} : \sum_{i \in S, a \in A} c^l(i,a) x_{ia} = \alpha^l\}.$$

For any $\{l_1, l_2, \ldots, l_h\} \subset \{1,2,\ldots,k\}$, let

$$\mathbb{Q}_{\{l_1,\ldots,l_h\}} := \{(x_{ia}) | ((x_{ia}), \delta) \in \hat{\mathbb{Q}}_{\{l_1,l_2,\ldots,l_h\}} \text{ for some } \delta \in \mathbb{R}^n\},$$
$$\text{where } \hat{\mathbb{Q}}_{\{l_1,\ldots,l_h\}} := \{((x_{ia}), \delta) \in \hat{\mathbb{Q}}^k : \mathcal{I}((x_{ia})) = \{l_1, l_2, \ldots, l_h\}\}.$$

For any compact convex set D we denote by $\text{ext}(D)$ the set of extreme points of D.

Then, we have the following, whose proof is done in Section 5.

Lemma 3.1. *Under Assumption* $(*)$*, it holds that for any* $\{l_1, \ldots, l_h\} \subset \{1, \ldots, k\}$,

$$\text{ext}\left(\mathbb{Q}_{\{l_1,\ldots,l_h\}}\right) \subset \{x(\beta, w, \delta) : (w, \delta) \in \Pi_S^k \times \mathcal{S}_S^k\}, \tag{3.6}$$

where k *is the number of constraints.*

The existence of a Pareto optimal pair of stationary policy and stopping time requiring randomization in at most k states is given in the following.

Theorem 3.2. *Suppose Assumption* $(*)$ *holds. Then a Pareto optimal pair* (π^*, τ^*) *for* **VCOP** *exists in* $\Pi_S^k \times \mathcal{S}_S^k$*, that is,*

$$\boldsymbol{e}(\{\mathbb{E}_\beta^\pi \boldsymbol{r}(x_\tau) | (\pi, \tau) \in \Lambda^k(\alpha, \beta)\}) \subset \boldsymbol{e}(\{\mathbb{E}_\beta^\pi \boldsymbol{r}(x_\delta) | (w, \delta) \in (\Pi_S^k \times \mathcal{S}_S^k) \cap \Lambda^k(\alpha, \beta)\}). \tag{3.7}$$

Proof. Let $(\pi^*, \tau^*) \in \boldsymbol{e}(\{\mathbb{E}_\beta^\pi \boldsymbol{r}(x_\tau) | (\pi, \tau) \in \Lambda^k(\alpha, \beta)\})$. Then, by Theorem 2.1, there sexists $b^* \in K^*$ such that (π^*, τ^*) is a optimal solution for (2.5) with $b = b^*$. Here, from Theorem 3.1 and Corollary 3.2, we can assume that $(\pi^*, \tau^*) \in \Pi_S \times \mathcal{S}_S$, so that there exists $\{l_1, l_2, \ldots, l_h\} \subset \{1,2,\ldots h\}$ with $x(\beta, \pi^*, \tau^*) \in \mathbb{Q}_{\{l_1,l_2,\ldots,l_h\}}$. Applying Lemma 3.1, there exists $(w^*, \tau^*) \in \Pi_S^k \times \mathcal{S}_S^k$ which is a solution of (2.5) with $b = b^*$. This means that (w^*, τ^*) is a Pareto optimal for **VCOP**, as required. ∎

Example 3.1

Consider the following numerical example with $p = 1$.
$S = \{1, 2, 3, 4\}, A = \{1\}, (\alpha_1, \alpha_2) = (0.5, 0.4), \beta = (0.25, 0.25, 0.25, 0.25)$,

$$(p_{ij}(1)) = \begin{pmatrix} 0.3 & 0.4 & 0.1 & 0.2 \\ 0.4 & 0.1 & 0.2 & 0.3 \\ 0.2 & 0.3 & 0.4 & 0.1 \\ 0.3 & 0.3 & 0.1 & 0.3 \end{pmatrix},$$

$$\begin{array}{cccc} c^1(1,1) = 0.6, & c^1(2,1) = 0.1, & c^1(3.1) = 0.5, & c^1(4,1) = 0.4, \\ c^2(1,1) = 0.6, & c^2(2,1) = 0.05, & c^2(3.1) = 0.1, & c^2(4,1) = 0.8, \\ r(1) = 4, & r(2) = 3, & r(3) = 2, & r(4) = 2. \end{array}$$

Letting $x_i = x_{i1}$ $(i \in S)$, the mathematical programming problem (**MMP(II)**) for the corresponding constrained optimization problem is given as follows:

$$\begin{array}{ll} \text{Maximize} & -x_1 - 0.1x_2 + 0.7x_3 + 0.9x_4 + 2.75 \\ \text{subject to} & x_1 = (0.25 + 0.3x_1 + 0.4x_2 + 0.2x_3 + 0.3x_4)(1 - \delta(1)), \\ & x_2 = (0.25 + 0.4x_1 + 0.1x_2 + 0.3x_3 + 0.3x_4)(1 - \delta(2)), \\ & x_3 = (0.25 + 0.1x_1 + 0.2x_2 + 0.4x_3 + 0.1x_4)(1 - \delta(3)), \\ & x_4 = (0.25 + 0.2x_1 + 0.3x_2 + 0.1x_3 + 0.3x_4)(1 - \delta(4)), \\ & 0.6x_1 + 0.1x_2 + 0.5x_3 + 0.4x_4 \leqq 0.5, \\ & 0.6x_1 + 0.05x_2 + 0.1x_3 + 0.8x_4 \leqq 0.4, \\ & x_i \geqq 0, 0 \leqq \delta(i) \leqq 1, i = 1, 2, 3, 4. \end{array}$$

After a simple calculation, we find the optimal solution of the above is $x_1^* = 0, x_2^* = 26/71$, $x_3^* = 43/71, x_4^* = 57/142$, $\delta^*(1) = 1, \delta^*(2) = 79/209, \delta^*(3) = 0, \delta^*(4) = 33/128$ and the optimal value is $1242/355 (\fallingdotseq 3.49859)$. Note that the value is $285/82 (\fallingdotseq 3.47561)$ for $\delta(1) = \delta(2) = 1$ and $\delta(3) = \delta(4) = 0$.

Thus, by Corollary 3.2, the pair $(w^*, \tau^*) \in \Pi_S^2 \times \mathcal{S}_S^2$ with $w^*(i) = 1$ for all $i \in S$ and $f^{\tau^*}(1) = \delta^*(1) = 1, f^{\tau^*}(2) = \delta^*(2) = 79/209, f^{\tau^*}(3) = \delta^*(3) = 0, f^{\tau^*}(4) = \delta^*(4) = 33/128$ is optimal for the corresponding constrained optimization problem and the optimal reward $1242/355$. Note that $\tau^* \in \mathcal{S}_S^2$.

4 Lagrange multiplier approaches

In this section, we define the Lagrangian associated with **VCOP** and the saddle-point statement is given(cf. [13]). Consequently, by solving a parametric Mathematical Programming problem defined in the sequel, a Pareto optimal pair is obtained.

Let $b = (b_1, \ldots, b_p) \in K^*$. The Lagrangian, L^b, associated with **VCOP** is defined as

$$L^b((\pi, \tau), \lambda) := \sum_{i=1}^{p} b_i \overline{\mathbb{E}}_\beta^\pi(r^i(X_\tau)) + \sum_{l=1}^{k} \lambda_l(\alpha^l - \overline{\mathbb{E}}_\beta^\pi(\sum_{t=1}^{\tau-1} c^l(X_t, \Delta_t))) \tag{4.1}$$

for any $(\pi, \tau) \in \Pi \times \mathcal{S}$ and $\lambda = (\lambda_1, \ldots, \lambda_k) \in \mathbb{R}_+^k$, where $\mathbb{R}_+^k$ is the positive orthant of $\mathbb{R}^k$.

Hereafter $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_k) \in \mathbb{R}_+^k$ will be written simply by $\lambda \geqq 0$.

For the Lagrangian approach we shall refer to ([14]). We have the following saddle-point statement, whose proof is similar to (Theorem 2, p.221 in [14]) combined with the use of the scalarization technique and omitted.

Theorem 4.1. (cf. [14]) *For some $b \in K^*$, suppose that the Lagrangian L^b has a saddle-point at $(\pi^*, \tau^*) \in \Pi \times \mathcal{S}$ and $\lambda^* \in \mathbb{R}^k_+$, i.e.,*

$$L^b((\pi, \tau), \lambda^*) \leqq L^b((\pi^*, \tau^*), \lambda^*) \leqq L^b((\pi^*, \tau^*), \lambda) \tag{4.2}$$

for all $(\pi, \tau) \in \Pi \times \mathcal{S}$ and $\lambda \in \mathbb{R}^k_+$. Then, (π^*, τ^*) is a Pareto optimal for **VCOP**.

In order to have the existence of a saddle-point of the Lagrangian $L^b(b \in K^*)$ we introduce the set of $N \times K$ matrices as follows:
For $M > 0$, let

$$\mathbb{Q}(M) := \left\{ \begin{array}{l} (x_{ia}) \in \mathbb{R}^N \times \mathbb{R}^K : \\ \text{(i)} \ \sum_{a \in A} x_{ia} = \beta(i)(1 - \delta(i)) + \sum_{j \in S, a \in A} x_{ja} p_{ji}(a)(1 - \delta(i)) \ (i \in S) \\ \text{(ii)} \ 0 \leqq \delta(i) \leqq 1 \ \ (i \in S) \\ \text{(iii)} \ \sum_{i \in S, a \in A} x_{ia} \leqq M - 1 \\ \text{(iv)} \ x_{ia} \geqq 0 \ \ (i \in S, a \in A) \end{array} \right\}. \tag{4.3}$$

Note that $\mathbb{Q}(M)$ is identical with the set of feasible solutions of the Mathematical Programming problem introduced in [9] to solve stopped MDPs with a constrained stopping time and condition (iii) of (4.3) means $\overrightarrow{\mathbb{E}}^w_\beta \tau^\delta \leqq M$, where $w \in \Pi_S$ is constructed from (x_{ia}) through (3.4). Under Assumption $(*)$, it clearly holds that for a sufficient large $M > 0$

$$\mathbb{Q}^k \subset \mathbb{Q}(M). \tag{4.4}$$

Henceforth, $M > 0$ will be fixed such that (4.4) holds.

By using occupation measures defined in Section 2, the Lagrangian $L^b(b \in K^*)$ can be rewritten as follows:

$$L^b((x_{ia}), \lambda) := \sum_{i \in S} \sum_{l=1}^{p} b_l r^l(i) y_i + \sum_{l=1}^{k} \lambda_l (\alpha^l - \sum_{j \in S, a \in A} c^l(j, a) x_{ja}) \tag{4.5}$$

$$= \sum_{i \in S, a \in A} (\sum_{j \in S} p_{ij}(a) r^b(j) - r^b(i) - \sum_{l=1}^{k} \lambda_l c^l(i, a)) x_{ia} \tag{4.6}$$

$$+ \sum_{l=1}^{k} \lambda_l \alpha^l + \sum_{i \in S} r^b(i) \beta(i),$$

where $y_i := \beta(i) + \sum_{j \in S, a \in A} x_{ja} p_{ji}(a) - \sum_{a \in A} x_{ia}$ and $r^b(j) := \sum_{l=1}^k b_l r^l(j)$, for $(x_{ia}) \in \mathbb{Q}(M)$ and $\lambda \in \mathbb{R}^k_+$.

We need the following condition.

Assumption (**). (*Slater condition*) *There exists* $(x_{ia}) \in \mathbb{Q}(M)$ *such that*

$$\sum_{i \in S, a \in A} c^l(i,a) x_{ia} < \alpha^l \tag{4.7}$$

for all $l = 1, \ldots, k$.

Then, applying (Theorem 1, p.217 in [14]) we have the following Lemma under Slater condition.

Lemma 4.1. *Under Assumption* (*) *and* (**), *for any* $b \in K^*$, *the Lagrangian* L^b *has a saddle-point at* $(x^*_{ia}) \in \mathbb{Q}(M)$ *and* $\lambda^* \in \mathbb{R}^k_+$, *i.e.,*

$$L^b((x_{ia}), \lambda^*) \leqq L^b((x^*_{ia}), \lambda^*) \leqq L^b((x^*_{ia}), \lambda)$$

for all $(x_{ia}) \in \mathbb{Q}(M), \lambda \in \mathbb{R}^k_+$.

If we construct a stationary policy w^* from $(x^*_{ia}) \in \mathbb{Q}(M)$ in Lemma 4.1 through (3.4), (w^*, λ^*) satisfies (4.2). Thus, we have the following from Lemma 4.1.

Corollary 4.1. *Under Assumption* (*) *and* (**), *for any* $b \in K^*$, *the Lagrangian* $L^b(\cdot, \cdot)$ *has a saddle-point* $(w^*, \lambda^*) \in \Pi_S \times \mathbb{R}^k_+$.

Applying the results above, we can present a parametric Mathematical Programming approach to obtain a Pareto optimal pair for **VCOP**. For any $b \in K^*$ and $\lambda \in \mathbb{R}^k_+$, let

$$r(i,a|b,\lambda) := \sum_{j \in S} p_{ij}(a) r^b(j) - r^b(i) - \sum_{l=1}^{k} \lambda_l c^l(i,a). \tag{4.8}$$

For $b \in K^*$ and $\lambda \in \mathbb{R}^k_+$, a parametric Mathematical Programming problem $\mathbf{MP}(b, \lambda)$ will be given as follows:

$$\mathbf{MP}(b,\lambda): \text{ Maximize } \sum_{i \in S, a \in A} r(i,a|b,\lambda) x_{ia}$$
$$\text{subject to } (x_{ia}) \in \mathbb{Q}(M).$$

Then, by using a result in [9], for each $\lambda \geqq 0$ we have the optimal value $v(b,\lambda)$ for $\mathbf{MP}(b,\lambda)$. By (4.6) and Lemma 4.1, there exists $\lambda^* \in \mathbb{R}^k_+$ such that

$$v(b,\lambda^*) + \sum_{l=1}^{k} \lambda^*_l \alpha^l = \min_{\lambda \geqq 0} (v(b,\lambda) + \sum_{l=1}^{k} \lambda_l \alpha^l). \tag{4.9}$$

From this multiplier λ^*, we solve $\mathbf{MP}(b,\lambda^*)$. Let $((x^*_{ia}), \delta^*)$ be a solution of $\mathbf{MP}(b,\lambda^*)$. Then, from the discussion above, (w^*, δ^*) and λ^* is a saddle-point satisfying (4.2), and we can say that (w^*, δ^*) is a Pareto optimal pair for **VCOP** and the value of $\mathbf{MP}(b,\lambda^*)$ is the expected rewards corresponding the Pareto optimal pair (w^*, δ^*), where w^* is a stationary policy determined by x^*_{ia} through (3.4).

Example 4.1

This is Example 3.1. By solving (4.9) with $b = 1$, we get $\lambda^* = (29/213, 248/213)$ and the value of the saddle-point is 1242/355. In order to obtain a optimal pair for **VCOP**, we solve $\mathbf{MP}(1,\lambda^*)$ and get the optimal pair $(w^*, \tau^*) \in \Pi^2_S \times \mathcal{S}^2_S$ as follows: $w^*(i) = 1$ for all $i \in S$ and $f^{\tau^*}(1) = \delta^*(1) = 1, f^{\tau^*}(2) = \delta^*(2) = 79/209, f^{\tau^*}(3) = \delta^*(3) = 0, f^{\tau^*}(4) = \delta^*(4) = 33/128$ and the corresponding optimal reward 1242/355, which is equal to the numerical results in Example 3.1.

5 Proof of Lemma 3.1

In this section, we prove Lemma 3.1.

By argument similar to those used in (Theorem 3.8, p.34, in [1]) we can show that

$$\text{ext}\left(\mathbb{Q}_{\{l_1,\ldots,l_h\}}\right) \subset \{x(\beta,w,\delta) : (w,\delta) \in \Pi_S^k \times \mathcal{S}_S\}. \tag{5.1}$$

Let $(w^*,\delta^*) \in \Pi_S^k \times \mathcal{S}_S$ be such that $x(\beta,w^*,\delta^*) \in \mathbb{Q}_{\{l_1,\ldots,l_h\}}$. Suppose that there exists $j_n (n = 1,\ldots,h+1)$ with

$$0 < \delta^*(j_n) < 1 \quad \text{for} \quad n = 1,2,\ldots h+1. \tag{5.2}$$

For simplicity, put $x^* = x(\beta,w^*,\delta^*)$ suppressing β, w^* and δ^*.

Let

$$\begin{aligned} &L := \{l_1,l_2,\ldots,l_h\}, && && \overline{L} := \{1,2,\ldots,k\} - L, \\ &J := \{j_1,j_2,\ldots,j_{h+1}\} && \text{and} && \overline{J} := S - J. \end{aligned}$$

For any row vector $x = (x_1,x_2,\ldots,x_N) \in \mathbb{R}^n$, we can write $x = (x_J, x_{\overline{J}})$, where x_J and $x_{\overline{J}}$ are subvectors of x and $x_J = \{x_i : i \in J\}$ and $x_{\overline{J}} = \{x_i : i \in \overline{J}\}$. Also, $P^\delta(w^*)$ will be partitioned into submatrices as follows:

$$P^\delta(w^*) = \begin{pmatrix} P^\delta(w^*)_{JJ} & P^\delta(w^*)_{J\overline{J}} \\ P^\delta(w^*)_{\overline{J}J} & P^\delta(w^*)_{\overline{J}\overline{J}} \end{pmatrix},$$

where $P^\delta(w^*)_{JJ} = (P_{ij}(w^*)(1-\delta(j))), i \in J; j \in J$ and other submatrices are similarly defined.

For simplicity, we write

$$P^\delta(w^*) = \begin{pmatrix} P_1 & P_2 \\ P_3 & Q \end{pmatrix}.$$

Let $c(w^*) = (c_{ij}(w^*))$, where $c_{ij}(w^*) = c^j(i,w^*)$ for $i \in S$ and $j \in \{1,2,\ldots,k\}$. $C(w^*)$ will be partitioned as done in the above:

$$C(w^*) = \begin{pmatrix} C_{JL} & C_{J\overline{L}} \\ C_{\overline{J}L} & C_{\overline{J}\overline{L}} \end{pmatrix},$$

suppressing w^*.

Here we consider the following inequality system (cf. (3.5)).

$$\begin{aligned} &\text{(i)} && x_J = \beta_J(1-\delta_J) + x_J P_1 + x_{\overline{J}} P_3, \\ &\text{(ii)} && x_{\overline{J}} = \beta_{\overline{J}}(1-\delta_{\overline{J}}) + x_J P_2 + x_{\overline{J}} Q, \\ &\text{(iii)} && x_J C_{JL} + x_{\overline{J}} C_{\overline{J}L} = \alpha_L, \\ &\text{(iv)} && x_J C_{J\overline{L}} + x_{\overline{J}} C_{\overline{J}\overline{L}} < \alpha_{\overline{L}}, \end{aligned} \tag{5.3}$$

where $\beta_J(1-\delta_J) = (\beta(i)(1-\delta(i)); i \in J), \beta_{\overline{J}}(1-\delta_{\overline{J}}) = (\beta(i)(1-\delta(i)); i \in \overline{J})$ and $=$ and $<$ mean componentwise relations.

We note that $x^* = (x_J^*, x_{\overline{J}}^*)$ and $\delta^* = (\delta_J^*, \delta_{\overline{J}}^*)$ satisfy (5.3) obviously.

From Assumption (∗), it clearly holds that $\lim_{n\to\infty} Q^n = \mathbf{0}$, so that $(I-Q)^{-1}$ exists and by (ii) in (5.3) we get

$$x_{\overline{J}} = (\beta_{\overline{J}}(1-\delta_{\overline{J}}) + x_J P_2)(I-Q)^{-1}, \tag{5.4}$$

where I is an identity matrix with the same dimensions as Q.

Also, since (i) in (5.3) includes only δ_J with respect to δ, it uniquely determines δ_J if x_J and $\delta_{\overline{J}}$ are given. Thus (i) and (ii) in (5.3) determine uniquely $x_{\overline{J}}$ and δ_J if x_J and $\delta_{\overline{J}}$ are given. Inserting from (5.4) into (iii) in (5.3), we have that

$$x_J(C_{JL} + P_2(I-Q)^{-1}) = \alpha_L - \beta_{\overline{J}}(1-\delta_{\overline{J}})(I-Q)^{-1}c_{\overline{J}L}. \tag{5.5}$$

Now, we denote by $\hat{D}$ the set of all pairs $(x_J, \delta_{\overline{J}})$ satisfying (5.3).

Let D be the set of all $x_J, (x_J \geqq 0)$ satisfying (5.5) with $\delta_{\overline{J}} = \delta^*_{\overline{J}}$, that is,

$$D = \{x_J | (x_J, \delta^*_{\overline{J}}) \in \hat{D} \text{ and } x_J \geqq 0\}. \tag{5.6}$$

Observing that (5.5) with $\delta_{\overline{J}} = \delta^*_{\overline{J}}$ has h equations and $h+1$ unknown elements, we find that D is a polyhedral convex set with at least one dimension. Since (5.2) means that $x^*_J \in D$ is a relative interior point in D, there exists $0 < \gamma < 1$ and $x^1_J, x^2_J \in D$ with

$$x^*_J = \gamma x^1_J + (1-\gamma)x^2_J. \tag{5.7}$$

Let $x^1_{\overline{J}}, \delta^1_J$ and $x^2_{\overline{J}}, \delta^2_J$ be those determined uniquely thorough (i)–(ii) in (5.3) with $x_J = x^1_J, \delta_{\overline{J}} = \delta^*_{\overline{J}}$ and $x_J = x^2_J, \delta_{\overline{J}} = \delta^*_{\overline{J}}$ respectively. Let $x^1 = (x^1_J, x^1_{\overline{J}}), x^2 = (x^2_J, x^2_{\overline{J}}), \delta^1 = (\delta^1_J, \delta^*_{\overline{J}})$ and $\delta^2 = (\delta^2_J, \delta^*_{\overline{J}})$. We can assume that x^1 and x^2 satisfying (iv) in (5.3) by choosing x^1_J and x^2_J sufficiently near to x^*_J. Applying Lemma 2.1 we get

$$x^1 = x(\beta, w^*, \delta^1) \text{ and } x^2 = x(\beta, w^*, \delta^2).$$

Thus, we have that

$$x(\beta, w^*, \delta^*) = \gamma x(\beta, w^*, \delta^1) + (1-\gamma)x(\beta, w^*, \delta^2),$$

which implies $x(\beta, w^*, \delta^*) \notin \text{ext}(\mathbb{Q}_{\{l_1, l_2, \ldots, l_h\}})$. This completes the proof. ∎

References

[1] E. Altman. *Constrained Markov Decision Processes.* Chapmann & Hall/CRC, 1999.

[2] D. Assaf and E. Samuel-Cahn. Optimal multivariate stopping rules. *J. Appl. Probab.*, 35:693–706, 1998.

[3] H. P. Benson. An improved definition of proper efficiency for vector maximization with respect to cones. *J. Math. Anal. Appl.*, 71:232–241, 1979.

[4] V. S. Borkar. *Topics in controlled Markov chains.* Longman Scientific & Technical, Harlow, 1991.

[5] Y. S. Chow, H. Robbins, and D. Siegmund. *Great expectations: the theory of optimal stopping.* Houghton Mifflin Co., Boston, Mass., 1976.

[6] N. Furukawa. Characterization of optimal policies in vector-valued markovian decision processes. *Math. Oper. Res.*, 5:271–279, 1980.

[7] N. Furukawa and S. Iwamoto. Stopped decision processes on complete separable metric spaces. *J. Math. Anal. Appl.*, 31:615–658, 1970.

[8] A. Hordijk. *Dynamic programming and Markov potential theory.* Mathematical Centre Tracts, No. 51. Mathematisch Centrum, Amsterdam, 1974.

[9] M. Horiguchi. Markov decision processes with a stopping time constraint. *To appear in Mathematical Methods of Operations Research*, 53:issue 2, 2001.

[10] M. Horiguchi, M. Kurano, and M. Yasuda. Markov decision process with constrained stopping times. *Proceedings of 39th IEEE Conference on Decision and Control(CDC2000). Markov Decision Processes*, 1:706–710, 2000.

[11] A. Irle. Minimax results and randomization for certain stochastic games. *In: Ricceri B, Stephen S (eds.)Minimax Theory and Applications.*, pages 91–103, 1998.

[12] D. P. Kennedy. On a constrained optimal stopping problem. *J. Appl. Probab.*, 19:631–641, 1982.

[13] M. Kurano, J. Nakagami, and Y. Huang. Constrained markov decision processes with compact state and action spaces: The average case. *Optimization*, 48:255–269, 2000.

[14] D. G. Luenberger. *Optimization by vector space methods.* John Wiley & Sons, Inc., New York, 1969.

[15] D. C. Nachman. Optimal stopping with a horizon constraint. *Math. Oper. Res.*, 5:126–134, 1980.

[16] Ulrich Rieder. On stopped decision processes with discrete time parameter. *Stochastic Processes Appl.*, 3(4):365–383, 1975.

[17] J. Stoer and C. Witzgall. *Convexity and Optimization in Finite Dimensions I.* Springer-Verlag, Berlin and Newyork, 1970.

[18] K. Wakuta. Optimal stationary policies in the vector-valued markov decision processes. *Stochastic Processes and their Applications*, 42:149–156, 1992.